Knowledge Engineering and Data Science (KEDS)
Vol 2, No 1, Juni 2019, pp. 41–46

pISSN 2597-4602
eISSN 2597-4637

Adam Optimization Algorithm for Wide and Deep Neural Network

Imran Khan Mohd Jais, Amelia Ritahani Ismail *, Syed Qamrun Nisa

Department of Computer Science, Kulliyyah of Information and Communication Technology
International Islamic University Malaysia, P.O. Box 10, 50728 Kuala Lumpur, Malaysia

*amelia@iium.edu.my**
** corresponding author*

ARTICLE INFO

Article history:
Received 4 March 2019
Revised 6 April 2019
Accepted 19 May 2019
Published online 23 June 2019

Keywords:
wide and deep network
neural network
adam algorithm
breast cancer dataset

ABSTRACT

The objective of this research is to evaluate the effects of Adam when used together with a wide and deep neural network. The dataset used was a diagnostic breast cancer dataset taken from UCI Machine Learning. Then, the dataset was fed into a conventional neural network for a benchmark test. Afterwards, the dataset was fed into the wide and deep neural network with and without Adam. It was found that there were improvements in the result of the wide and deep network with Adam. In conclusion, Adam is able to improve the performance of a wide and deep neural network.

I. Introduction

Neural networks are often used to solve classification and recommendation problems. This research will work on classifying whether a tumor is malignant or benign. However, the main objective of this research will focus on the effects of Adam on the performance of the wide and deep network. A challenge in working with conventional neural networks is to achieve both memorization and generalization. According to [1] :

> *"Memorization can be loosely defined as learning the frequent co-occurrence of items or features and exploiting the correlation available in the historical data. Generalization, on the other hand, is based on transitivity of correlation and explores new feature combinations that have never or rarely occurred in the past."*

Next, it was found that neural networks with high number of features have a tendency to over-generalize and give irrelevant outputs [1]. Due to this, there is a high tendency to get false predictions or misleading results.

The wide and deep network combines the benefits of memorization and generalization by adding [illegible] models and deep neural networks and it is highly useful for large scale regression and classification problems.

In 2018, [2] investigated a Convolutional Neural Networks (CNNs) based computer-aided diagnosis (CAD) framework for breast cancer classification. In general, deep learning may require extensive datasets to organize systems while transfer learning method consumes a little datasets of medical images. Transfer learning method optimize the training of The CNNs. As a result, the CNN achieved the finest outcomes with 98.94% of accuracy.

In 2017, [3] proposed CNNs to classify the hematoxylin and eosin stained breast biopsy images. The designed network architecture retrieved different scales information such as nuclei and overall tissue organization. This design extend the proposed system to whole-slide histology images. Furthermore, the CNNs extracted features are also used to train a SVM based classification engine.

https://doi.org/10.17977/um017v2i12019p41-46

The use of CAD systems increases the diagnosis efficiency as well as the level of inter-observer agreement.

On the other hand, [1] investigate on how wide linear models can effectively memorize sparse feature interactions using cross-product feature transformations, while deep neural networks can simplify to formerly unseen feature interactions through low dimensional embeddings. Online experiment results show that wide & deep model significantly increased app acquisitions compared with wide-only and deep-only models.

Furthermore, [4] analyse the theoretical convergence of the algorithm properties and deliver a regret bound on the convergence rate that is comparable to the best known results under the online convex optimisation framework. Empirical results demonstrate that Adam works well in practice and compares favorably to other stochastic optimization methods. Stochastic gradient descent is an efficient and effective optimization approach that was central in many machine learning success stories. One example of it is a novel advances in deep learning.

In 2016, [5] proposed a wide and deep neural network with strong induction ability to model the transformation, and an efficient training strategy. The promising approach has potential application in image based ophthalmologic diseases diagnosis. It may provide a fresh, general, high-performance computing framework for image segmentation. Moreover, [6] introduced ReQuIK, a multi-perspective query suggestion system for children. The system provides the suggestion process by applying a wide and deep neural network ranking strategy that considers both raw text and traits, generally associated with kid-related queries. By applying a multi-perspective approach based on deep learning, the proposed query suggestion module is able to learn distinctive characteristics that portray adults and children queries.

The application of deep learning has in recent years lead to a dramatic boost in performance in many areas such as computer vision, speech recognition or natural language processing [7]. Despite this huge empirical success, the theoretical understanding of deep learning is still limited. In this paper we address the non-convex optimization problem of training a feedforward neural network. This problem turns out to be very difficult as there can be exponentially many distinct local minima [8]. It has been shown that the training of a network with a single neuron with a variety of activation functions turns out to be NP-hard.

II. Method

A. Data Collection

The dataset was taken from UCI Machine Learning and was titled “Breast Cancer Wisconsin” [9]. The dataset consisted of 3 categories which are mean, standard error, and worst. Then, each category contains 10 features which are radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. In total, there are 30 features.

B. Data Preprocessing

The Figure 1 shows one example of the distribution of each feature in the original dataset. Due to the uneven distribution of the data, we use a normalization technique called Z-Score to transform the distribution of the data into a more uniform manner. The Figure 2 shows the distribution of the data after normalization.

From the Figure 1 and Figure 2, we can see that there is a drastic change in the distribution of the data after normalization. The distribution among features in each category are more even and has low variance which is a good thing for the machine learning algorithm so it can learn better.

C. Feature Correlation

This research is done with Python using the *Tensorflow* library which was developed by *Google* [10]. Figure 3 shows a heatmap of the correlation strength between features. As we can see from the bar on the right, a lighter color represents a stronger correlation between the features. From this figure, we can see a patch of strongly correlated features on the bottom left. From there it was shown that radius_mean, perimeter_mean, and area_mean is highly correlated with area_worst, perimeter_worst, and radius_worst.

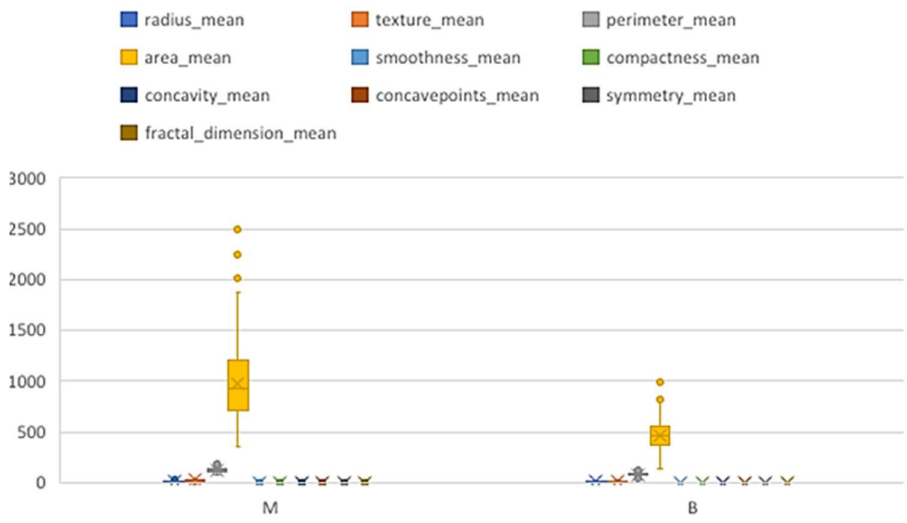

Fig. 1. Distribution of ten features in the dataset

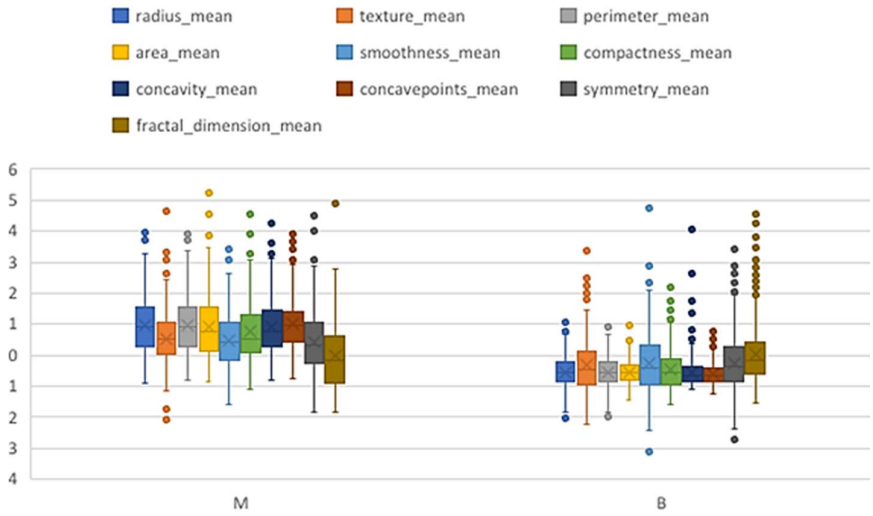

Fig. 2. The data distribution after normalization

Furthermore, it was also shown on the top left that radius_mean, perimeter_mean, and area_mean have strong correlation between themselves. On the other hand, we can see that there are a few weakly correlated features across the heatmap but since it also has strong correlation with other features, we decided to not prune any feature.

D. Machine Learning

The wide and deep neural network requires the user to define which features are base features, crossed features, and deep features. The purpose of this is to define which features go into the wide and deep part of the network. In this case, since the features are already grouped into 3 parts, the process is simplified. The mean group is the base features, the standard error group is the crossed features, and the last group is the deep features. Other parameters that had to be defined are shown in the Table 1.

Moreover, we also record the time taken for the model to complete training. Next, the model is run for the benchmark test. For the benchmark test, all features are fed into the deep part of the network with the same parameters as above. Afterwards, the model was run using the wide and deep network without Adam optimization then with Adam optimization.

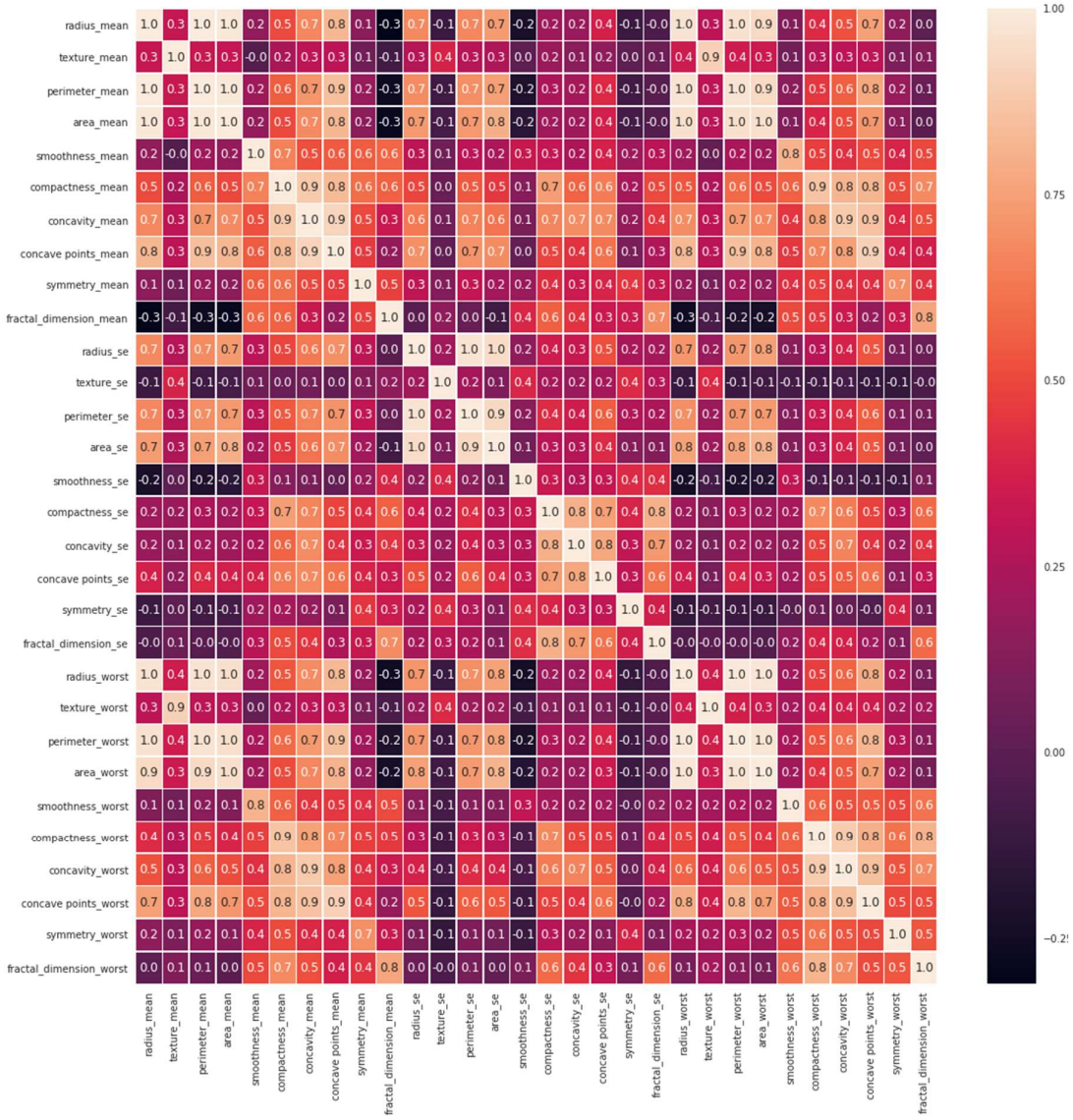

Fig. 3. Correlation strength between features

Table 1. Parameter detail

Paramater	Description
Training set	455 rows
Test set	114 rows
Number of examples per batch	30
Number of training epochs	50
Number of hidden layers	6
Number of neurons for each layer	100, 75, 50, 25, 10, 5

III. Results and Discussion

Table 2 shows the results obtained after training has completed. Based on the results, the first thing to notice is that wide and deep network with Adam optimization obtained the highest accuracy. However, we could not automatically assume that it is the best performer because the high accuracy result could be due to overfitting. Thus, we need to consider the next two metrics.

Table 2. Training results

Parameter	Neural network	Wide and deep neural network without Adam optimization	Wide and deep neural network with Adam optimization
Accuracy	0.965	0.947	0.991
AUC	0.993	0.995	0.996
AUC precision recall	0.989	0.990	0.993
Average loss	0.123	0.165	0.137
Loss	3.704	4.705	3.896
Prediction/mean	0.301	0.292	0.431
Time taken	29.959	55.627	51.647

AUC refers to the area under the curve of the Receiver Operating Characteristic (ROC) line while the next metric refers to the area under the precision recall curve. The ROC curve shows the true positive rate against the false positive rate for the test set. On the other hand, the precision recall graph shows the precision rate against the recall rate as the name suggests. Precision refers to the number of true positives over the total of true positives and false positives while recall refers to the number of true positives over the total of true positives and false negatives.

For both of these metrics, the area under the curve needs to be close to 1 to show that the result obtained is good and was not due to overfitting. In this case, for the wide and deep network with Adam optimization, the number is close to 1 so we can be assured that the result obtained was not due to any error or anomaly.

The average loss and loss metric refers to the loss function which is a part of the learning process in a neural network. As we can see both neural network and wide and deep network with Adam optimization resulted in a low average loss and loss which is a good sign.

The prediction/mean shows a close value to the label/mean which is 0.333 and that is also signs of a good model because it means that the prediction rate is close to the truth. Next, we can see improvement in the time taken for the wide and deep network with Adam optimization to finish training compared to without optimization. However, the conventional neural network showed the shortest time but this could be due to the network architecture being more simple compared to the wide and deep network.

IV. Conclusion

As a conclusion, it was expected to see that the wide and deep network with Adam optimization performs the best. However, the wide and deep network without Adam optimization trails not far from it. In this case, we may need to scale the model bigger to see significant difference between their performances by feeding a bigger dataset as an example. Nonetheless, it was proven that Adam optimization is able to improve the performance of the wide and deep neural network.

Acknowledgment

This research was supported by the Research Initiative Grants Scheme (RIGS): RIGS16-346-0510.

References

[1] H. Cheng *et al.*, "Wide & Deep Learning for Recommender Systems," in *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, 2016, pp. 7–10.

[2] H. Chougrad and H. Zouaki, "Deep Convolutional Neural Networks for Breast Cancer Screening," *Comput. Methods Programs Biomed.*, vol. 157, pp. 19–30, 2018.

[3] T. Araújo *et al.*, "Classification of breast cancer histology images using convolutional neural networks," *PLoS One*, vol. 12, no. 6, 2017.

[4] D. P. Kingma and J. L. Ba, "ADAM: a method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, 2015, pp. 1–15.

[5] Q. Li, B. Feng, L. Xie, P. Liang, H. Zhang, and T. Wang, "A Cross-Modality Learning Approach for Vessel Segmentation in Retinal Images," *IEEE Trans. Med. Imaging*, vol. 35, no. 1, pp. 109–118.

[6] I. M. Azpiazu, N. Dragovic, O. Anuyah, and M. S. Pera, "Looking for the Movie Seven or Sven from the Movie Frozen ? A Multi-perspective Strategy for Recommending Queries for Children," in *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, 2018, pp. 92–101.

[7] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A Survey of Deep Neural Network Architectures and Their Applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.

[8] I. Safran and O. Shamir, "On the Quality of the Initial Basin in Overspecified Neural Networks," in *International Conference on Machine Learning*, 2016, vol. 774–782.

[9] W. H. Wolberg and O. L. Street, W. Nick Mangasarian, "Breast Cancer Wisconsin (Diagnostic) Data Set," 1995. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

[10] Google, "Tensorflow," *2018*. [Online]. Available: Https://Github.Com/Tensorflow.